

TFW #H



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/765,883	01/29/2004	Hironori Yasukawa	500.43450X00	2307

24956 7590 11/30/2005
MATTINGLY, STANGER, MALUR & BRUNDIDGE, P.C.
1800 DIAGONAL ROAD
SUITE 370
ALEXANDRIA, VA 22314

EXAMINER
SAVLA, ARPAN P.

ART UNIT 2185
PAPER NUMBER

DATE MAILED: 11/30/2005

Please find below and/or attached an Office communication concerning this application or proceeding.

RECEIVED
OIPE/IAP
DEC 01 2005

Best Available Copy

Office Action Summary	Application No. 10/765,883	Applicant(s) YASUKAWA ET AL.	
	Examiner Arpan P. Savla	Art Unit 2185	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 29 January 2004.
- 2a) ☐ This action is **FINAL**. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) 1-7 is/are pending in the application.
- 4a) Of the above claim(s) _____ is/are withdrawn from consideration.
- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1-7 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☒ The specification is objected to by the Examiner.
- 10) ☒ The drawing(s) filed on 29 January 2004 is/are: a) ☒ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☒ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☒ All b) ☐ Some * c) ☐ None of:
1. ☒ Certified copies of the priority documents have been received.
2. ☐ Certified copies of the priority documents have been received in Application No. _____.
3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application (PTO-152) |
| 3) <input checked="" type="checkbox"/> Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)
Paper No(s)/Mail Date <u>1/29/04 & 6/14/05</u> | 6) <input type="checkbox"/> Other: _____ |

DETAILED ACTION

The instant application having Application No. 10/765883 has a total of 7 claims pending in the application, there are 3 independent claims and 4 dependent claims, all of which are ready for examination by the examiner.

INFORMATION CONCERNING OATH/DECLARATION

Oath/Declaration

1. Applicant's oath/declaration has been reviewed by Examiner and is found to conform to the requirements prescribed in 37 CFR 1.63.

STATUS OF CLAIM FOR PRIORITY IN THE APPLICATION

2. As required by MPEP § 201.14(c), acknowledgment is made of Applicant's claim for priority based on an application filed in the Japanese Patent Office on November 28, 2003.

INFORMATION CONCERNING DRAWINGS

Drawings

3. Applicant's drawings submitted are acceptable for examination purposes.

Art Unit: 2185

ACKNOWLEDGMENT OF REFERENCES CITED BY APPLICANT

Information Disclosure Statement

4. As required by MPEP § 609(c), Applicant's submission of the Information Disclosure Statements dated January 29, 2004 and June 14, 2005 are acknowledged by Examiner and cited references have been considered in the examination of the claims now pending. As required by MPEP § 609 c(2), a copy of the PTOL-1449 initialed and dated by Examiner is attached to the instant office action.

OBJECTIONS

Specification

5. The title of the invention is not descriptive. A new title is required that is clearly indicative of the invention to which the claims are directed.

The following title is suggested: "System And Method For A Storage Control Apparatus Using Information On Management Of Storage Resources".

Claims

6. **Claims 1, 3-4, and 6-7** are objected to because of the following informalities:

7. **As per claims 1, 4, and 7**, the phrase "data control I/O unit" in claims 1 and 4 and the phrases "channel control unit" and "disk control unit" in claim 7 are inconsistent with the specification. Examiner believes that said phrases refer to "data control I/O section", "channel control section", and "disk control section" respectively from the

Art Unit: 2185

specification. Applicant must be consistent throughout both the specification and claims and choose either "section" or "unit" to describe the claimed inventions.

8. **As per claims 3 and 6**, the phrase "an RAID" in line 3 of claim 3 and line 4 of claim 6 should read "a RAID".

9. **Also for claim 3**, the phrase "include of a" in line 2 should read "include a".

10. **As per claim 4**, on line 11 there should be a semicolon after the word "request".

Examiner also suggests that line 12 and 14 of claim 4 be indented to clearly signify new limitations within the claim.

Appropriate corrections are required.

REJECTIONS NOT BASED ON PRIOR ART

Claim Rejections - 35 USC § 112

11. The following is a quotation of the second paragraph of 35 U.S.C. 112:

The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

12. **Claims 1, 4, and 7 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.** The phrase "can be communicatively connected" in line 3 of claims 1, 4, and 7 does not clearly identify if the "plurality of communication ports" are actually communicatively connected to the "plurality of information processing apparatuses" or not.

REJECTIONS BASED ON PRIOR ART

Claim Rejections - 35 USC § 103

13. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

14. **Claims 1-7 are rejected under 35 U.S.C. 103(a) as being obvious over Blumenau et al. (U.S. Patent 6,260,120) in view of Voigt et al. (U.S. Patent 5,960,451).**

15. **As per claim 1**, Blumenau discloses a storage control apparatus comprising:
a data I/O control unit which has a plurality of communication ports that can be communicatively connected with any of a plurality of information processing apparatuses (col. 8, lines 24-28, 36-37, and 40-41; col. 9, lines 50-57; Fig. 1, elements 20, 21, 22-25, 27, and 35-36; and Fig. 2, elements 41-44), is communicatively connected to a plurality of physical disk drives for storing data (col. 8, lines 24-35, 36-37, and 41-44; and Fig. 1, elements 20, 26, 28-31, and 37-38), receives a data I/O request for data stored in the physical disk drives from the information processing apparatuses via the communication ports (col. 8, lines 48-49), and performs data read/write from/to the physical disk drives in accordance with the received data I/O request (col. 8, lines 56-60); It should be noted that "cached storage subsystem" is analogous to "storage control apparatus", "storage controller" is analogous to "data I/O control unit", and "hosts" are analogous to "information processing apparatuses".

a first memory storing a data which is read/written among the data stored in the physical disk drives (col. 8, lines 36-37, 48-54, and 60-65; and Fig. 1, element 32); and

a second memory storing information on management of storage resources including the communication ports and the physical disk drives (col. 27, lines 23-33 and Fig. 25, element 282); It should be noted that “virtual ports” are analogous to “communication ports”. It should also be noted that the logical storage volumes directly correspond to the storage devices (i.e. physical disk drives), see col. 8, lines 28-29.

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface, an identifier of the communication port and an identifier of the physical disk drive are transmitted to said user interface (col. 30, line 59 – col. 31, line 2 and Fig. 30, elements 346 and 347). It should be noted that “clicking on it with a pointing device” is analogous to “transmission request” and “system administrator” is analogous to “user”. Also, see citation note directly above regarding logical storage volumes.

Blumenau does not disclose expressly a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses;

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface a storage capacity of the first memory which have been allocated for said user are transmitted to said user interface.

Voigt discloses a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses (col. 6, lines 13-16; col. 7, lines 30-31; and Fig. 2, element 56);

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface a storage capacity of the first memory which have been allocated for said user are transmitted to said user interface (col. 6, lines 13-17; col. 7, lines 5-9 and 30-36; Fig. 4, elements 90, 104, and 106). It should be noted that "as the administrator moves the sliding bar" acts as a "transmission request".

Blumenau and Voigt are analogous art because they are from the same field of endeavor, that being storage systems that use logical storage units (LUNs) with graphical user interfaces (GUIs).

At the time of the invention it would have been obvious to a person of ordinary skill in the art to combine Voigt's LUN cache storage capacity indicator and GUI with Blumenau's cached storage subsystem and GUI.

The motivation for doing so would have been because in a system with fixed physical capacity, it would be beneficial to determine how much usable capacity can be afforded simultaneously for each logical unit type, given the diversity of consumption rates among the various types (Voigt, col. 3, lines 15-19).

Therefore, it would have been obvious to combine Blumenau and Voigt for the benefit of obtaining the invention as specified in claim 1.

Art Unit: 2185

16. **As per claims 2 and 5**, Blumenau discloses information on management of the storage resources includes:

first correlation between the physical disk drive and a data amount which can be stored in the first memory among the data stored in the physical disk drive (col. 8, lines 56-62); It should be noted that when taking the broadest interpretation of the claim language it is clear that the limitations of the claim do not identify what "correlation" specifically entails or the size of the "data amount". Blumenau discloses reading data from the storage devices and writing the data (amount not specified, but nonetheless still a discrete amount of data) back to cache memory, thus disclosing a correlation between the storage devices and cache memory.

and information representing a second correlation between the first correlation and the communication port (col. 8, lines 62-65). Again, it should be noted that when taking the broadest interpretation of the claim language it is clear that the limitations of the claim do not identify what "correlation" specifically entails. Blumenau discloses that the data used in the "first correlation" (see citation directly above) is written to the cache memory by the port adapters (which contain at least two ports, see col. 9, lines 54-55), thus disclosing a correlation between the first correlation and the communication port.

17. **As per claims 3 and 6**, Blumenau discloses physical disk drives include a plurality of hard disk drives constituting a RAID (col. 9, lines 16-19).

18. **As per claim 4**, Blumenau discloses a storage control apparatus comprising: a data I/O control unit which has a plurality of communication ports that can be communicatively connected with any of a plurality of information processing

Art Unit: 2185

apparatuses (col. 8, lines 24-28, 36-37, and 40-41; col. 9, lines 50-57; Fig. 1, elements 20, 21, 22-25, 27, and 35-36; and Fig. 2, elements 41-44), is communicatively connected to a plurality of physical disk drives for storing data (col. 8, lines 24-35, 36-37, and 41-44; and Fig. 1, elements 20, 26, 28-31, and 37-38), receives a data I/O request for data stored in the physical disk drives from the information processing apparatuses via the communication ports (col. 8, lines 48-49), and performs data read/write from/to the physical disk drives in accordance with the received data I/O request (col. 8, lines 56-60); It should be noted that "cached storage subsystem" is analogous to "storage control apparatus", "storage controller" is analogous to "data I/O control unit", and "hosts" are analogous to "information processing apparatuses".

a first memory storing a data which is read/written among the data stored in the physical disk drives (col. 8, lines 36-37, 48-54, and 60-65; and Fig. 1, element 32); and

a second memory storing information on management of storage resources including the communication ports and the physical disk drives (col. 27, lines 23-33 and Fig. 25, element 282); It should be noted that "virtual ports" are analogous to "communication ports". It should also be noted that the logical storage volumes directly correspond to the storage devices (i.e. physical disk drives), see col. 8, lines 28-29.

said method comprising the steps of:

receiving a transmission request of the information on management of the storage resource from a user via a user interface (col. 30, lines 59-62). It should be noted that "clicking on it with a pointing device" is analogous to "transmission request" and "system administrator" is analogous to "user".

and in response to said receiving step, transmitting an identifier of the communication port and an identifier of the physical disk drive (col. 30, line 62 – col. 31, line 2 and Fig. 30, elements 346 and 347). Also, see citation note directly above regarding logical storage volumes.

Blumenau does not disclose expressly a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses;

in response to said receiving step, transmitting a storage capacity of the first memory which have been allocated for said user to said user interface.

Voigt discloses a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses (col. 6, lines 13-16; col. 7, lines 30-31; and Fig. 2, element 56);

in response to said receiving step, transmitting a storage capacity of the first memory which have been allocated for said user to said user interface (col. 6, lines 13-17; col. 7, lines 5-9 and 30-36; Fig. 4, elements 90, 104, and 106). It should be noted that “as the administrator moves the sliding bar” acts as a “transmission request”.

19. **As per claim 7**, Blumenau discloses a storage control apparatus comprising:

a channel control unit which has a plurality of communication ports that can be communicatively connected with any of a plurality of information processing apparatuses and receives a data I/O request for data stored in physical disk drives including a plurality of hard disk drives constituting an RAID (col. 8, lines 24-28, 36-37,

40-41, and 48-49; col. 9, lines 16-19 and 50-57; Fig. 1, elements 20, 21, 22-25, 27, and 35-36; and Fig. 2, elements 41-44); It should be noted that “cached storage subsystem” is analogous to “storage control apparatus”, “port adapter” is analogous to “channel control unit”, and “hosts” are analogous to “information processing apparatuses”.

a disk control unit which is communicatively connected to the physical disk drives and performs data read/write from/to the physical disk drives according to the data I/O request (col. 8, lines 24-35, 36-37, 41-44, 56-60; and Fig. 1, elements 20, 26, 28-31, and 37-38); It should be noted that “storage adapter” is analogous to “disk control unit”.

a first memory storing a data which is read/written among the data stored in the physical disk drives (col. 8, lines 36-37, 48-54, and 60-65; and Fig. 1, element 32); and

a second memory storing information on management of storage resources including the communication ports and the physical disk drives (col. 27, lines 23-33 and Fig. 25, element 282); It should be noted that “virtual ports” are analogous to “communication ports”. It should also be noted that the logical storage volumes directly correspond to the storage devices (i.e. physical disk drives), see col. 8, lines 28-29.

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface, an identifier of the communication port and an identifier of the physical disk drive are transmitted to said user interface (col. 30, line 59 – col. 31, line 2 and Fig. 30, elements 346 and 347). It should be noted that “clicking on it with a pointing device” is analogous to “transmission request” and “system administrator” is analogous to “user”. Also, see citation note directly above regarding logical storage volumes.

Blumenau does not disclose expressly a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses;

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface a storage capacity of the first memory which have been allocated for said user are transmitted to said user interface.

Voigt discloses a second memory storing information on management of storage resources including a storage capacity of the first memory allocated for each user using the information processing apparatuses (col. 6, lines 13-16; col. 7, lines 30-31; and Fig. 2, element 56);

wherein in response to reception of a transmission request of the information on management of the storage resource from a user via a user interface a storage capacity of the first memory which have been allocated for said user are transmitted to said user interface (col. 6, lines 13-17; col. 7, lines 5-9 and 30-36; Fig. 4, elements 90, 104, and 106). It should be noted that "as the administrator moves the sliding bar" acts as a "transmission request".

RELEVANT ART CITED BY THE EXAMINER

The following prior art made of record and not relied upon is cited to establish the level of skill in Applicant's art and those arts considered reasonably pertinent to Applicant's disclosure. See MPEP 707.05(e).

The following reference discloses a **storage area network (SAN) comprised of a RAID array**.

U.S. Patent Application Publication Number

2003/0093501

Conclusion

STATUS OF CLAIMS IN THE APPLICATION

The following is a summary of the treatment and status of all claims in the application as recommended by MPEP 707.70(i):

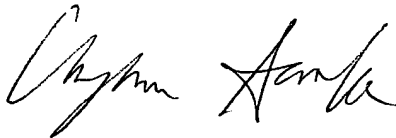
CLAIMS REJECTED IN THE APPLICATION

Per the instant office action, claims 1-7 have received a first action on the merits and are subject of a first action non-final.

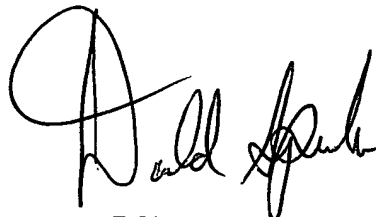
Any inquiry concerning this communication or earlier communications from the examiner should be directed to Arpan P. Savla whose telephone number is (571) 272-1077. The examiner can normally be reached on M-F 8:30-5.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Donald Sparks can be reached on (571) 272-4201. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).



Arpan Savla
Assistant Examiner
11/17/05



DONALD SPARKS
SUPERVISORY PATENT EXAMINER

**FORM PTO-1449 U.S. Department of
Commerce Patent and Trademark Office**

ATTY. DOCKET NO.

SERIAL NO.

500.43450X00

10/765,883

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**

(Use several sheets if necessary)

APPLICANT
H. YASUKAWA, et alFILING DATE
January 29, 2004

GROUP 2185


U.S. PATENT DOCUMENTS

EXAMINER INITIAL	DOCUMENT NUMBER	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
A.S.	4 4 6 7 4 2 1	8/84	White	—	—	
A.S.	5 2 1 0 8 4 4	5/93	Shimura et al	—	—	
A.S.	6 0 1 4 7 3 0	1/00	Ohtsu	—	—	
A.S.	6 4 2 1 7 1 1	7/02	Blumenau et al	—	—	
A.S.	6 4 8 0 9 3 2	11/02	Vallis et al	—	—	
A.S.	2 0 0 1 0 8 4 3	1/02	Sanada et al	—	—	

FOREIGN PATENT DOCUMENTS

	DOCUMENT NUMBER	DATE	COUNTRY	CLASS	SUBCLASS	ABSTRACT	
						YES	NO
A.S.	1 3 5 7 4 7 6	10/03	European	—	—		
A.S.	0 1 9 5 1 1 3	12/01	PCT	—	—		

OTHER DOCUMENTS (Including Author, Title, Date, Pertinent Pages, Etc.)

EXAMINER

DATE CONSIDERED

11/4/05

EXAMINER: Initial if citation is considered, draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

(Form PTO-1449 [6-4])

FORM PTO-1449 U.S. Department of
Commerce Patent and Trademark Office

ATTY. DOCKET NO.

SERIAL NO.

500.43450X00

10/765,883

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**

(Use several sheets if necessary)

APPLICANT
H. YASUKAWA, et al

FILING DATE
January 29, 2004

GROUP 2185

U.S. PATENT DOCUMENTS

EXAMINER INITIAL	DOCUMENT NUMBER	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
A.S.	2 0 1 3 3 6 6 9	9/02	Devireddy et al			
A.S.	2 0 1 5 6 9 8 4	10/02	Padovano			
A.S.	2 0 1 2 3 0 6 8	6/04	Hashimoto			
A.S.	2 0 0 5 0 2 6 8	3/05	Yoshida			

FOREIGN PATENT DOCUMENTS

DOCUMENT NUMBER	DATE	COUNTRY	CLASS	SUBCLASS	ABSTRACT
					YES NO

OTHER DOCUMENTS (Including Author, Title, Date, Pertinent Pages, Etc.)

EXAMINER

DATE CONSIDERED

EXAMINER: Initial if citation is considered, draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

(Form PTO-1449 [6-4])

INFORMATION UNDER 37 CFR 1.56(a)

(For Initial Filing)

The following references are submitted as information
to comply with the duty of disclosure under 37 CFR 1.56(a):

References	Disclosed in the specification?		Copy			Translation	
	Yes	No	Enc.	Follow	Please obtain	Enc.	Not available
AS 1. JP-A-5-128002	<input type="radio"/>		<input type="radio"/>			<input type="radio"/> (only abstract)	
2.							
3.							
4.							
5.							
6.							
7.							
8.							
9.							
10.							

By: [Signature]

Date Considered: 11/4/05

Notice of References Cited	Application/Control No. 10/765,883		Applicant(s)/Patent Under Reexamination YASUKAWA ET AL.	
	Examiner Arpan P. Savla		Art Unit 2185	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-6,260,120	07-2001	Blumenau et al.	711/152
*	B	US-5,960,451	09-1999	Voigt et al.	711/114
*	C	US-2003/0093501	05-2003	Carlson et al.	709/220
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Peter Zabback et al., "The RAID Configuration Tool", 19-22 Dec. 1996, 3rd International Conference on High Performance Computing, Proceedings, pp. 55-61.
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

The RAID Configuration Tool

Peter Zabback
Tandem Computers
10100 N. Tantau Ave.
Cupertino, CA 95014
zabback@patch.tandem.com

Jai Menon
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120
menonjm@almaden.ibm.com

Jeff Riegel
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120
riegel@almaden.ibm.com

Abstract

Disk Arrays or RAID's are a widely accepted I/O system architecture, useful for a wide range of applications. Before a RAID can be used, one needs to configure the RAID to select parameters such as the RAID level to use, the stripe unit to use, how large of a cache to use, and so on. Selecting these configuration parameters can be quite complex, yet no aids are available today to help the user configure his RAID. The optimal selection of the parameters strongly depends on the specific workload characteristics of the application. In this paper, we describe a configuration tool called Raidtool which is intended to support the systems designer in the selection of the configuration parameters. Our approach consists of three basic steps. The first step is to collect a trace of I/Os while running one or more typical applications. In the second step, this trace data is analyzed to determine the workload characteristics of the applications. In the third and final step, we use a simulator to evaluate the different RAID controller configurations.

1. Introduction

Disk Arrays or RAID's have become a widely accepted I/O architecture. Even relatively cheap single card RAID controllers allow for a wide range of configuration parameters such as RAID level, size of stripe unit, size of optional caches, etc., yet no aids are available today to help the user configure his RAID. The optimal parameter settings strongly depend on the characteristics of the applications I/O-workload. In this paper we describe a software tool called Raidtool which supports the system designer in the process of configuring a RAID controller.

Our main design goal was to come up with an easy to use and fast software tool which allows the efficient evaluation of different configuration alternatives.

The first step in the configuration process is to collect workload information by means of I/O-traces. In a second

step these traces are analyzed to identify the key workload characteristics of the application. The third step is the simulation of different configuration alternatives. These simulations are driven by a synthetic workload generator, which generates a workload according to the characteristics determined in the second step.

The rest of this paper describes our three step approach in some detail and is organized as follows: in Section 2 we discuss the features of a disk array controller modeled by Raidtool. Section 3 describes our overall approach; section 4 describes the components of the workload description provided by the analyzer to the simulator. In section 5 we discuss the simulator in some detail. Finally, in section 6 we show some results from our tool compared to a fully trace-driven approach.

2. RAID Architecture Model

Raidtool accepts a description of the RAID controller card as its input. In order for our tool to support a wide range of different controller cards, we came up with a very generic and highly configurable controller card description. A typical RAID controller card contains a microprocessor. The card also contains a certain amount of DRAM for instructions and operands (Data Buffer in Figure 1). This DRAM is optimized for reading and writing small chunks of data (i.e., instructions and operands). In addition there is typically another DRAM for data caching. A RAID controller card may optionally be equipped with non-volatile RAM (NVRAM); while the regular volatile cache is used for read caching only, this NVRAM is used for write caching. All these components of a controller card are connected by an internal bus with a certain bandwidth. The characteristics (speed, bandwidth, size, etc.) of the components as well as the bus are configurable in our tool.

The RAID controller connects to a number of disks. For this controller-disk connection, Raidtool currently supports SCSI-2 [1] or SSA [2] connections. The performance characteristics of these connections are again configurable.

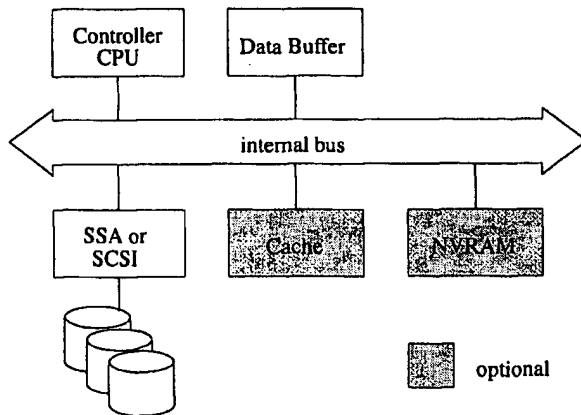


Figure 1. Principal Disk Array Architecture

2.1. RAID Features

In our model, a RAID controller card is configurable for a (unlimited) number of RAID groups. Each group may be configured for a different RAID level; we support all the usual RAID levels [5, 3]. Each group may consist of different disk types; within one group we assume the same disk type. Disks are configurable and also support track buffering. Examples of parameters supported for a RAID group are the RAID level, stripe unit, group size, etc. See our research report [7] for a complete list.

2.2. NVRAM

RAIDs with the optional NVRAM installed can do *fast write* [4]. The host is notified about the completion of a write request as soon as the data has arrived in the NVRAM cache. The physical write to disk (destage) is performed asynchronously at a later point in time.

Our RAID model assumes that when a configuration with NVRAM is chosen, all writes are fast writes. Writes that arrive when the NVRAM is full are delayed until space becomes available.

3. The Analyzer/Simulation Approach

3.1. Overview

Raidtool works in three main steps (Figure 2). The first step is to collect information about the typical workload in terms of I/O trace data. In the second step, this trace data is analyzed and a specific workload description is generated. This workload description is used to generate a synthetic

load with comparable characteristics. Finally, in the third step, we simulate different disk array controller configurations. This simulation is driven by a synthetic load, which is generated according to the workload characteristics determined by the trace analyzer. The simulator is run several times, to simulate various I/O arrival rates, including rates faster than the actual I/O rate in the original trace.

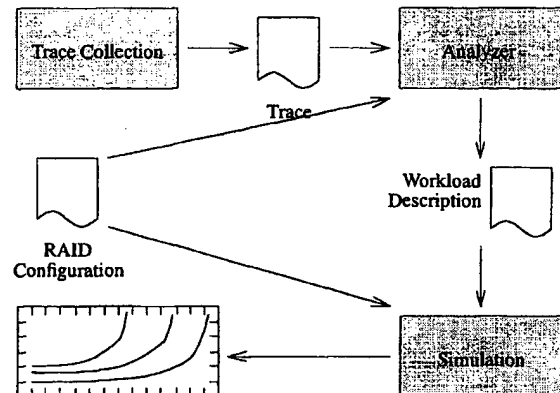


Figure 2. Analyzer/Simulation Approach

There are two main reasons for taking this analyzer/simulation approach instead of running the simulation directly from the I/O traces. First, the tool is intended for interactive use, so performance is a major concern. The analyzer needs to be run only once for a given configuration, whereas the simulator must be run several times with different arrival rates, so we improve performance by performing as much of the work as possible (e.g. cache modeling) in the analyzer. The second reason is that we were not satisfied that any of the known techniques for speeding up a given trace did an adequate job of preserving the dependencies of the original I/O requests. Speeding up a trace is required in order to perform simulation at I/O rates higher than that of the original trace.

Instead of the simulation, one could try to use an analytical model to evaluate the configuration alternatives. Even though an analytical model allows for very quick evaluations, this approach is not practical because of the inherent complexity of the system being modeled. One could try to simplify the model to make it more tractable, but most of these simplifications are hard to justify or may even invalidate the analytical approach.

3.2. The Analyzer

The analyzer basically takes three items of information: a description of the traced system, a description of the RAID controller configuration, and the trace data itself. Based on

this information, the analyzer identifies the key workload characteristics which are important for the configuration of a RAID controller.

3.3. The Simulator

Using the workload description from the analyzer, the simulator generates a synthetic workload and models the selected RAID configuration. The synthetic workload consists of a number of generated I/O requests which are submitted to the simulated RAID controller. These I/O requests are generated using exponential arrival times. The overall arrival rate must be specified by the user for each run of the simulator. The simulator uses the workload characteristics obtained from the output of the analyzer to generate the size and address of each I/O request.

4. The Workload Description

This section describes the elements of the workload description provided by the analyzer to the simulator.

4.1. Seek Distribution

The main workload characteristic provided by the analyzer is a set of distributions of logical seek distances, which allows us to keep track of the spatial locality of the I/O-requests. The analyzer concatenates the address spaces of the traced disks to one large logical address space, which allows us to determine the logical block address of an I/O request. A *logical seek* is the distance between the logical block addresses of consecutive I/O-requests in the trace.

The analyzer divides the logical address space into multiple address ranges, and generates a separate logical seek distribution for each address range, as shown in figure 3. This method was chosen because the alternatives of a simple average seek distance or a single overall seek distribution (independent of position) were found to be inadequate. We discuss this further in our research report [7].

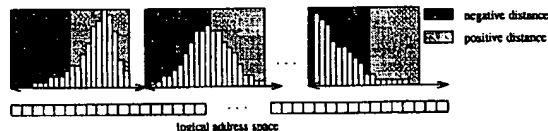


Figure 3. Multiple seek distributions, one per address range

4.2. Request Size Distribution

As with the logical seek information, providing only the average request size was deemed insufficient to capture the original distribution of request sizes. In addition, sizes for read and write requests may follow extremely different distributions. Thus, the analyzer generates a request size distribution for read and write requests separately. For each possible request size, the probability that a request has this size is given. From this request size/probability information the simulator calculates the Cumulative Distribution Function (CDF) of the request size distribution and generates pseudo random request sizes according to this distribution.

4.3. Interarrival Time Distribution

The next important workload characteristic is the distribution of the interarrival time. The simplest (and most commonly used) approach is to calculate the average interarrival time from the original trace and assume exponential interarrivals. Unfortunately, we observed that this assumption is not generally true. Figure 4 shows the arrival rate of requests as a function of evolving time for a real I/O trace. The overall average arrival rate for this trace is about 5.2 [req/s]. During most of the time the trace has an arrival rate close to 1 [req/s]; only during two peak periods is the arrival rate significantly higher. The overall observed maximum arrival rate is larger than 400 [req/s]. It is quite inappropriate to describe this behavior with an exponential interarrival time distribution whose mean is 0.19 [s]¹.

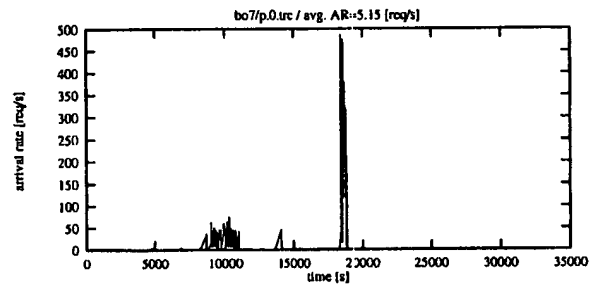


Figure 4. Arrival rate vs. evolving time for an I/O trace

The bottom line is that the assumption of exponential interarrival times is not always appropriate because some traces show a higher degree of "burstiness" than we can model with an exponential distribution.

¹the interarrival time of 0.19 [s] corresponds to an arrival rate of 5.2 [req/s].

4.4. Analyzing the Peak Periods of a Trace

Even though the assumption of exponential arrivals may not be appropriate for an entire trace, we may find time intervals in the trace which are better behaved in terms of their interarrival time distribution. In addition, the customer is only interested in the high load phases of a trace, since low load periods do not cause any performance problems. We therefore developed a simple filter, which takes a trace as input and separates the trace periods with high load.

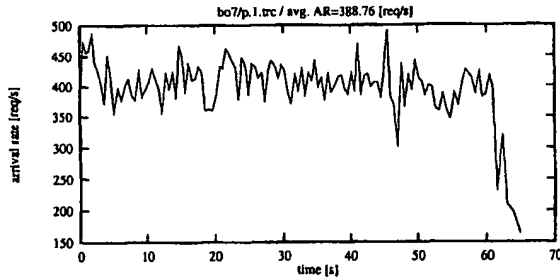


Figure 5. Arrival rate vs. evolving time for the first peak period of the AFS-trace

Figure 5 shows the average arrival rate as a function of evolving time for the first peak identified by the above described peak determination algorithm. The figure shows that we still observe a certain fluctuation in the average arrival rate, but the trace is much more well behaved with respect to the arrival rate, so it is reasonable to use an exponential distribution for the interarrival times. To account for minor variations in the arrival rate over time, the analyzer provides the simulator with a different arrival rate for each interval (where an interval is some fixed number of I/Os.)

4.5. Cache Hit Ratios

The analyzer generates hit ratios for the different RAM and NVRAM caches. We distinguish three types of cache hits:

- **Read hit:** A read request finds the requested data in the regular RAM cache or in the NVRAM cache.
- **Write hit on clean:** A write request finds a clean copy (value same as on disk) of the data to be updated in the RAM cache. This may save a read operation during later destage of the data from cache.
- **Write hit on dirty:** A write request finds the data to be updated in NVRAM cache. This old version is overwritten. This saves the destaging of the old version.

In addition, the analyzer generates hit ratios for the disk buffers. For each disk we get a separate buffer hit ratio.

Start Time	End Time	LBA	Size
0.000	1.000	1000	8
0.500	2.500	2000	64
0.750	2.000	3000	128
1.000	3.000	1008	8
2.000	4.000	3128	128
2.500	5.000	2064	64

Figure 6. A trace containing interleaved thread I/Os

4.6. Destaging

For configurations with NVRAM, all writes are submitted to the NVRAM and, at some later point in time, destaged to disk. Since destaging is part of the cache logic, it is modeled in the analyzer, which supplies the simulator with the distribution of the size and physical seek distance of destage I/O operations. Each destaging operation is a read or a write to one disk. Given the destage I/O information from the analyzer, the simulator adds destaging load to the disks. See [7] for a full description of the algorithm used.

4.7. Thread Lists

The logical seek distribution approach described in Section 4.1 produces fairly accurate results for most workloads. However, there are certain workloads which are not described well by this approach. These are workloads which consist of multiple series or "threads" of interleaved sequential I/Os. An example is shown in Figure 6.

The difficulty with sequential threads arises due to the fact that a pair of successive I/Os in a thread are separated by unrelated I/Os of another thread. This causes large logical seeks between the two I/Os of the thread. At such high seek values, there is not enough granularity in the seek distributions to enable the second I/O of the pair to return exactly to the point where the first left off.

To solve this problem, we decided to use a heuristic to explicitly identify threads of sequential I/Os. This heuristic identifies two I/Os as part of the same thread if the second's logical block address immediately follows the first's (i.e., the second I/O's LBA is equal to the first I/O's LBA plus the first I/O's request size in blocks.) There must be no more than a small number of I/Os intervening between the

two I/Os of the thread. Using this technique, the analyzer outputs multiple thread lists; each list describes the threads active during each interval of some number of I/Os. Each thread in the list is described by the number of I/Os in the thread, the starting block number, the total size of all the I/Os in the thread, etc. For each I/O it generates, the simulator determines whether the I/O is a thread I/O, and which thread it is a part of, based on the total number of thread I/Os that occurred during the interval.

In general, each sequential thread I/O in the original trace is dependent on its predecessor, and does not start until its predecessor has completed. To capture this behavior, the simulator keeps track of which threads are still active, i.e. have an I/O outstanding which has not yet completed. Only a thread that is not already active may be chosen for a newly-generated I/O to be a part of. If all threads are already active, the I/O is delayed until one or more threads become available.

The simulator simulates a given configuration at a variety of arrival rates. Increasing the arrival rate compared to the original trace will also change the distribution of thread I/Os. Thus, it is necessary for the simulator to modify the list of thread I/Os provided by the analyzer when a different arrival rate is used. To increase the arrival rate by x , we merge x adjacent lists of threads into one. This makes more threads available for the simulator to choose from at any given instant of time, similar to the effect that would occur from speeding up the trace.

5. RAID Controller Simulation

5.1. Cache Configurations

The simulator does no cache simulation, but determines randomly whether a given I/O request is a cache hit, with probability equal to the appropriate hit ratio provided by the analyzer. The simulator supports 4 basic cache configurations: no cache, read cache only, write cache only, and read/write cache.

5.2. I/O Scheduling

An I/O request generated by the simulator is converted to the appropriate disk-level I/Os based on the RAID configuration chosen for simulation. The following sections describe the timing of these disk-level I/Os.

5.2.1. Timing of Read I/Os

Figure 7 illustrates the timing of a read miss, as it is realized in the simulation program. Components of the read I/O response time are: $t_{CPU,ReadMiss}$, which denotes the CPU overhead for a read miss, and t_{Disk} , which denotes the disk

service time. The disk service time may be further broken up as shown, into the queuing delay, seek time, head settle time, rotational latency, transfer time, and controller overhead. The figure shows a read request which is serviced by five disks. All five I/O operations are issued at the same time and the "slowest" of these five disks determines the response time of the entire read request.

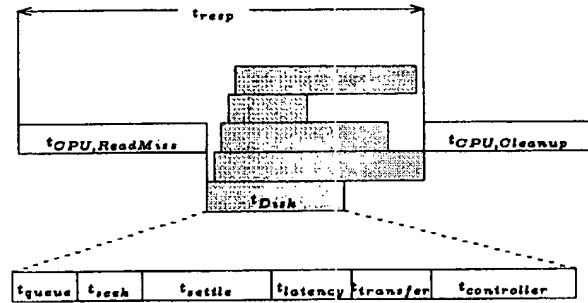


Figure 7. Components of read I/O response time t_{resp} (read miss)

5.2.2. Timing of Write I/Os

With NVRAM we assume that each write I/O is a fast write. The response time in this case is given by the time the I/O spent in the NVRAM queue waiting for sufficient NVRAM to become available plus CPU overhead for a fast write. Without NVRAM, each write is converted into four operations (in RAID 5): a read and a write operation at the data disk and a read and a write operation at the data disk [3].

In the simulation, we provide for the selection of one of three different scheduling strategies. These are: AlwaysSteal, in which we allow the scheduling of other requests, i.e. the "stealing" of the disk arm, between the read and write operations; StealOnParity, in which we only allow stealing of the disk arm between the read and write of parity; and NeverSteal, in which we disallow stealing of the disk arm between the read and write operations. See [7] for more information.

6. Results

To verify the validity and accuracy of our analyzer/simulation approach, we conducted a brief comparison study between our results and those of a traditional fully trace-driven simulation. For the role of a trace-driven simulation, we used the trace analyzer. Since it is a component of Raidtool, the RAID model it uses is identical to that of the full analyzer/simulation approach. It is able to produce

Trace	Dur. [s]	# Req	Arr Rate	Rd Frac [%]
A	177	49894	285	97
B	133	50000	375	43
C	167	50000	300	70
D	717	67774	95	20

Table 1. Traces for simulation

Conf	Cache	NVRAM	RAID Conf
1	0	0	11+P RAID-5
2	0	0	11+P RAID-0
3	0	0	2x(5+P RAID-5)
4	0	0	31+P RAID-5
5	0	0	31+P RAID-0
6	0	8 MB	31+P RAID-5
7	16 MB	0	31+P RAID-5
8	0	8 MB	11+P RAID-5
9	0	16 MB	11+P RAID-5

Table 2. Configurations used

response time values and performs full cache and NVRAM modeling.

We used four main traces for testing our tool, as shown in Table 1. Trace A was collected from a networked file server during a period of peak activity. This trace contained a very large number of sequential threads. Trace B was collected from a DB2 database server running the TPC-C benchmark. Trace C was generated using a synthetic trace generation program. Trace D was a trace of on-line transaction processing activity collected from a database server.

The various configurations we tested are shown in Table 2. Average response time results for both our tool and the trace-driven simulator are shown in Table 3. These results are at the original I/O rate of the trace.

These results show that our tool can approach quite closely to the trace-driven results at the original trace's I/O rate. This may be enough for some purposes. However, our analyzer already provides a value for the response time results at the original I/O rate, so the main purpose of our analyzer/simulation approach is to provide quick results at higher arrival rates.

Thus, we increased the arrival rate of the traces in order to compare the sped-up results to those of our tool for some of the configurations. To do this, we used the simple method of decreasing the interarrival time of each I/O by the factor that the trace is being sped up. This method was preferred over more complex methods such as folding [6] because the latter will greatly change some of the workload characteristics determined by the analyzer, such as the logical seek distributions and cache hit ratios.

Trace	Conf	RT (tool)	RT (trc-driv)
A	1	18.96	23.31
	2	13.97	17.05
	3	73.60	58.88
B	4	32.61	40.20
	5	13.22	13.22
	6	5.65	6.40
C	4	24.60	29.45
	7	19.08	20.51
D	1	48.14	46.27
	8	2.74	3.05
	9	2.85	2.97

Table 3. Simulation results (response times in milliseconds)

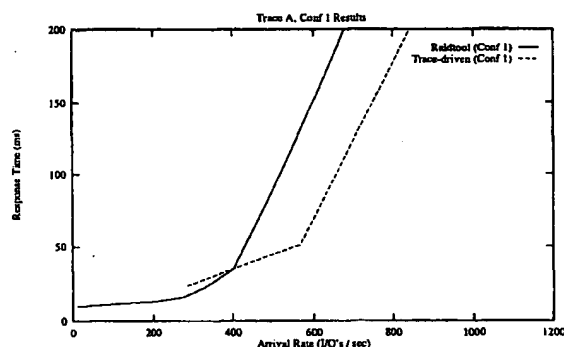


Figure 8. Simulation results for Trc A, Conf 1

Figure 8 shows results for Trace A at configuration 1. These results show that the two methods are quite close at arrival rates close to the trace, then begin to diverge at increased arrival rates. Figure 9 shows results for Trace A at configuration 3. Despite there being few writes in this trace, Conf. 3 (RAID-0) shows a significant advantage over Conf. 1 (RAID-5).

Figure 10 shows the results for Trace C at configurations 4 and 7. Note that although the tool does not match the trace-driven results exactly, it shows sufficiently the relative difference between these two configurations. It indicates clearly that the 16 MB of cache improves performance significantly.

Figure 11 shows the results for Trace D at configurations 1, 8, and 9. Here the curves are quite close and show the significant advantage gained by adding 8 MB of NVRAM (Conf 8) over none at all (Conf 1). However, increasing NVRAM to 16 MB (Conf 9) gains little improvement.

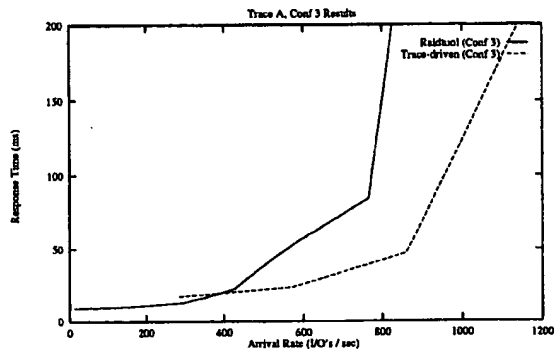


Figure 9. Simulation results for Trc A, Conf 3

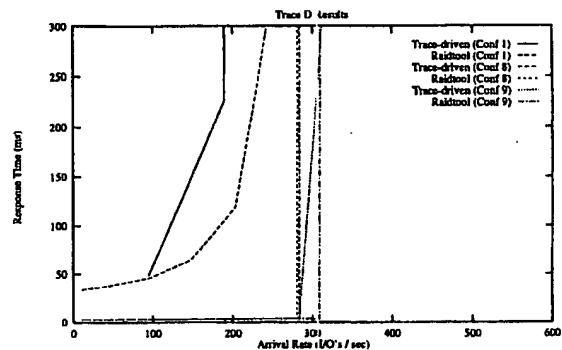


Figure 11. Simulation results for Trace D

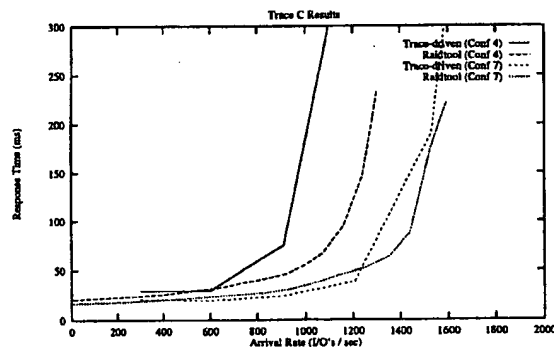


Figure 10. Simulation results for Trace C

7. Conclusion

Raidtool has been in use for approximately one year at IBM in a variety of situations. It consists of about 8500 lines of C++ code, and runs quickly enough to meet our initial goal for a fast, easy-to-use tool. It produces quite accurate approximations of results at arrival rates close to that of the original trace, which become somewhat more inaccurate as the arrival rate is increased. Nevertheless, the results appear to be accurate enough to allow a user to select the most appropriate RAID configuration.

References

- [1] American National Standards Institute. *Small Computer Systems Interface - 2, ANSI X3.131*, 1993.
- [2] American National Standards Institute. *Serial Storage Architecture SSA-PH (Transport Layer), ANSI X3T10.1*, 1995.
- [3] P. M. Chen, E. Lee, G. Gibson, D. Patterson, and R. Katz. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys*, 26(2):145-185, June 1994.
- [4] J. Menon. Performance of RAID 5 disk arrays with read and write caching. *Journal of Distributed and Parallel Databases*, 2(3):261-294, 1994.
- [5] D. Patterson, G. Gibson, and R. Katz. Reliable arrays of inexpensive disks (RAID). In *ACM Sigmod Conference 1988*, pages 109-116, Chicago, 1988.
- [6] K. Treiber and J. Menon. Simulation study of cached RAID5 designs. *IBM Research Report*, No. RJ 9823 (84886), 1994.
- [7] P. Zabback, J. Menon, and J. Riegel. The RAID configuration tool. *IBM Research Report*, publication in progress, 1996.

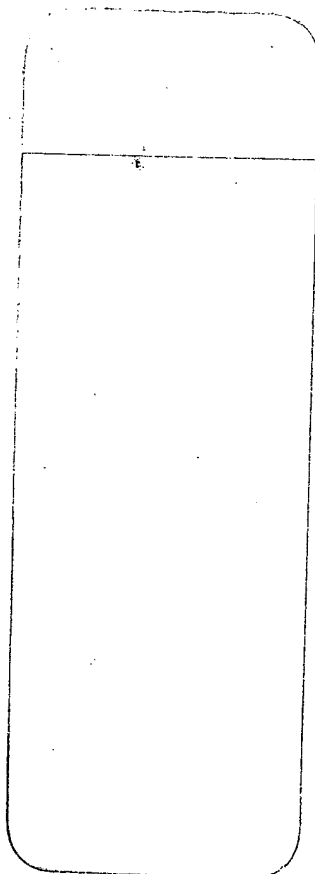
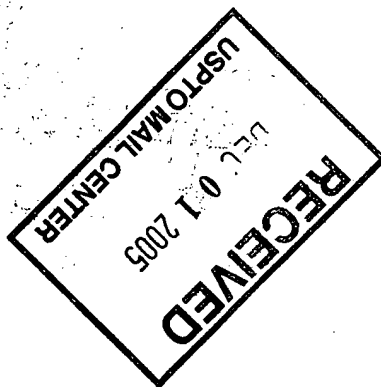
U.S. DEPARTMENT OF COMMERCE
COMMUNICATIONS CENTER FOR PATENTS
PO BOX 1450

ALEXANDRIA, VA 22313-1450

IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

AN EQUAL OPPORTUNITY EMPLOYER



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.